学校编码：10384　　分类号_____密级_____
学号：15420091151696　　UDC____

厦门大学

硕　士　学　位　论　文

信息不对称下的银行零售信用风险研究

——基于非参数随机森林

Research of Credit Risk under Asymmetric Information based on Nonparametric Random Forests

吴见彬

指导教师姓名：谢邦昌教授

专　业　名　称：统计学

论文提交日期：2012 年 4 月

论文答辩时间：2012 年 5 月

学位授予日期：

答辩委员会主席：_____

评　　阅　　人：_____

2012 年 4 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范(试行)》。

另外，该学位论文为(　　)课题(组)的研究成果，获得(　　)课题(组)经费或实验室的资助，在(　　)实验室完成。(请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。)

声明人(签名)：

2012 年 4 月 2 日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文(包括纸质版和电子版)，允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

(　　)1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

(　　)2.不保密，适用上述授权。

(请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。)

声明人(签名)：

2012 年 4 月 2 日

摘 要

近年来零售信贷业务发展迅猛，已经成为国内众多商业银行重点争夺的领域。与对公业务相比，零售业务呈现出客户量大但人均交易额小的特点，若使用人工逐户逐笔的方式进行审批和监管，不仅经营成本高业务效率低，还会出现人工评价标准不统一的问题。此外，银行零售市场具有信贷信息不对称的特征，这是信用风险产生的主要原因之一。在信贷信息不对称下，如何利用统计分析、数据挖掘等高新技术，建立可靠的分析模型，对用户的行为进行模型化自动化风险识别具有非常重要的意义。

本文首先剖析了随机森林的特征选择功能，指出随机森林计算出的变量重要性是有偏的，主要体现为：该重要性偏好于连续变量以及多类别的属性变量，同时还会受到输入变量间相关性的影响。本文在第 3 章中进行了实验模拟，模拟结果进一步验证了该变量重要性的有偏性。故本文引入了基于条件的森林 Cforest，可以计算出变量的条件重要性，被证明具有无偏性质。

在此基础上，本文分别对信用卡和个人住房贷款业务建立了随机森林违约识别模型，该模型具有很高的识别能力，优于基准的罗吉斯模型和 SVM 模型。此外，个人住房贷款由于合同周期长，数据常常呈现截尾性特征，故本文在第 6 章引入了针对截尾数据分析的随机生存森林(RSF)，它在识别违约与否的同时，还能计算出违约概率和违约时间，具有很大的应用价值。

本文的实证研究发现信用卡持有者的职业、学历、性别、婚姻、家庭人口数等借款人的个人特质在信用评估中最具影响力；而在个人住房贷款业务中，最具影响力的变量为贷款总额、借款人行业、贷款收入比和利率等，相比之下贷款状况变量比个人特质变量更重要。此外，在个人住房贷款业务中，违约率较低的客户群主要包括以下几类：高学历客户、中间年龄段客户、已婚客户，女性客户等。从职业来看，违约率最低的是公务员，其次为教师。

关键词：信息不对称；信用风险；随机森林；随机生存森林；Cforest

Abstract

The retail credit business is developing rapidly in recent years and has become the key field of many domestic commercial banks. Compared with the public business, retail business has much more customers but smaller amount per customer. In this case, using manual audit, not only leads to high operating costs and low efficiency, but also causes conflicting manual evaluation standards. In addition, the information of retail banking market is asymmetric between banks and customers, which is one of the main reasons for credit risk. Under this circumstance, it is of great importance to use statistical technology and data mining technology to establish a reliable analytical model.

This paper analyzed the feature selection function of random forest and pointed out that variable importance based on random forest (RF) is biased, as this importance prefers continuous variables and categorical variable with more categories. In addition, when input variables are correlated with each other, the importance is also biased. In Chapter 3, simulation results validate the bias of importance from RF. Based on this, we introduce condition based forest (Cforest) which can provide unbiased variable importance.

The empirical results showed that RF credit identify model is superior to SVM and logistic model. In addition, as individual housing loans are of long term and the data are often censored, this paper introduced random survival forests (RSF) to analyze the censored data which can not only recognize defaulters, but also predict the probability and time of default events.

The empirical result from credit card data shows that variables about borrowers' personal characteristics are more important in credit assessment, such as occupation, education, gender, marital status and family population. In contrast, empirical result from housing loan data shows that the most influential variables are amount of loan, industry of borrower, loan to income(LTI) and interest rates, which seemed variables about loan status is more important than variables about personal characteristics. This empirical result also shows that customers with following characters have lower default rates: well educated, medium-age, married and female. In addition, the lowest default rate happened among civil servants, followed by teachers.

Key Words: Information Asymmetry; Credit Risk; Random Forests; Random Survival Forests; Cforest

目 录

Contents

第 1 章 绪 论

1.1 研究背景

自 2000 年以来，我国商业银行的零售业务发展迅速，国内许多商业银行已经把零售业务发展作为主要发展战略。特别是对于新兴商业银行而言，零售业务是他们可以与四大国有商业银行公平竞争的领域。与对公业务不同，零售业务面向的是大量的个人客户，个人客户更注重的是银行的效率和服务，而这正是新兴商业银行的优势所在。也因此，零售业务成为了国内商业银行的重点争夺的领域。零售业务目前正朝着信用化的方向发展，特别是信用卡、个人住房贷款业务蓬勃发展，这就要求银行具有较高的信用风险管理水平。目前我国商业银行已经初步形成了以个人按揭、消费信贷、信用卡为主导的零售业务产品体系。但总的来说，由于国内银行起步较晚，信用风险管理依然比较落后。

与对公业务相比，银行零售业市场存在严重的信贷信息不对称现象，主要体现在商业银行很难掌握申请客户的全部信息。由于不良申请客户信息披露得不够充分可靠，特别是在国家个人社会信用系统，信用环境和法制环境不够完善的条件下，客户可能隐瞒自身真实收入和风险状况，致使商业银行对其缺乏足够的认识。信息不对称是信用风险产生的主要原因之一。不良客户恶意消费使用，造成商业银行的不良透支大量累积，运作效率低下，信息不对称面不断扩大，形成了一种有效信息的漏斗效应，风险随之增强，形成高风险问题客户“驱逐”低风险潜力客户的恶性循环。

在信贷信息不对称下，提高信用风险管理水平，一方面需要获得尽可能真实完善的客户信息，但在我国个人社会信用系统，信用环境和法制环境不够完善的条件下，获得完备可靠的信息不仅成本高而且几乎不可能；另一方面就是着重研究有效信用风险管理方法来弥补信息不对称的缺陷，比如国外银行都高度重视对历史数据的分析和利用，建立了高度整合的数据仓库，充分利用数据挖掘技术和各种统计分析，构建信用风险评估指标，提高信用风险的识别和管理能力。而我国各大银行由于数据积累不足或其他各方面的原因，各项管理决策还主要是依靠经验，存在很大的风险和隐患。

与其它业务相比，银行零售业务最大的特点是业务数量庞大，依靠经验进行

人工审批这会存在以下的问题：首先，零售业务数量庞大但客户分散，且每笔金额较小，使用人工审批成本高昂；其次，人工审批效率低，耗费时间长，在激烈的市场竞争中，审批耗时过长，可能导致一部分客户流向其它银行；再次，人工审批主观性过大，且难以制定统一标准化的决策。因此，对银行零售业信用风险进行自动化、模型化的管理是将来发展的趋势。因此，在信贷信息不对称下，如何利用统计分析、数据挖掘等高新技术研究，建立可靠的分析模型，对银行零售业客户的行为进行自动化智能化风险识别和预测，有效提高我国银行零售业市场的风险管理水平具有重要的理论和现实意义。

商业银行零售业务主要分为储蓄业务、消费信贷业务(个人住房贷款和汽车消费贷款)和信用卡业务。在这三项业务中，储蓄业务主要面临操作风险，而消费信贷业务和信用卡业务主要面临着信用风险。故本文以信用卡业务和个人住房贷款业务为例，建立随机森林信用风险评估模型。信用卡业务的最大问题是一些恶意违约的客户，以不良动机申请信用卡，进行信用卡欺诈，故在信用卡风险管理中，有效识别不良客户是重中之重。与信用卡业务不同，个人住房贷款业务一般不存在恶意欺诈的现象，违约产生的原因主要是由于失业或收入下降导致无力偿还；另一个不同之处是个人住房贷款采用分期付款的形式，因而借款人的违约时间会影响到银行的违约损失率。因此，对于个人住房贷款业务而言，计算违约的发生时间和概率也同样重要。针对这两种业务的不同特点，本文分别采用了随机分类森林和随机生存森林来建模，随机分类森林的优点是分类的准确性非常高，能够有效地识别出违约客户，而随机生存森林的优点是可以计算出违约时间和违约概率，但准确性低于随机分类森林。鉴于此，本文利用随机分类森林识别信用卡的违约客户，利用随机生存森林预测个人住房贷款客户的违约时间和违约概率。

目前随机森林已经应用十分广泛，但同时在国内外都出现了一些误用和滥用的情况，特别是在随机森林的特征选择部分，还有许多问题尚未解决，简单搬用随机森林的特征选择会对实证分析产生误导。因此，本文在国际最新研究成果的基础上，对随机森林的一些主要问题进行了详细地介绍和模拟，希望能为随机森林在信用风险管理领域的应用提供一个较为规范的分析思路和框架，为今后该领域的研究提供一定的参考。

1.2 信用风险及研究现状

信用风险又称违约风险，是指受信人不能履行还本付息的责任而使授信人的预期收益与实际收益发生偏离的可能性。一般来说，信用风险可分为理性违约风险和被迫违约风险。理性违约是指借款人通过对还款进行成本收益分析，认为违约更有利可图，例如房价大幅下跌，使得该房产的价值低于未偿还贷款总额，因而理性的借款人会选择放弃该房产而发生违约；被迫违约是指借款人由于失业、收入下降等原因，无力偿还贷款而产生的违约。

信用风险度量方法主要有专家法、评级法和信用评分法。专家法是根据专家的经验而建立的一套信用评分方法，根据借款人的 5 项因素(包括品德与声望、偿付能力、资金实力、担保、经营条件和商业周期，简称 5C)进行判断，做出贷款决定。评级法是将银行贷款分成若干等级，不同的等级赋予不同的损失准备金率，然后计算损失准备金并加总，就得出银行需要准备的用于防范风险的资本。

信用评分方法是以评价对象的相关指标为解释变量，运用数理统计方法建立模型，以模型输出的信用分值或违约概率与基准值比较，度量评价对象的风险大小(郭英见、吴冲，2009)。常用的信用评分模型有距离判别分析、罗吉斯模型、神经网络、决策树、支持向量机等模型。信用评分方法的应用最为有效，已是国际学术界和金融实业界研究信用风险的主流方法。

本文重点研究的是信用卡市场和个人住房贷款市场，故在此分别对这两类市场的研究现状进行总结归纳。

1.2.1 信用卡研究现状

信用卡的信用风险是指持卡人不能或不愿按照信贷协议约定偿还本息，从而对银行经营造成伤害的可能性。信用风险是信用卡业务面临的主要风险。及时有效地应对可能发生的信用风险，不论从商业银行自身而言，还是从监管机构而言，都对信用卡风险的度量突出了很高的要求。

很多学者研究了使用信用卡的影响因素，对于信用卡的信用风险管理具有一定的指导意义，比如 Gross 等(2002)认为客户居住地、年龄、收入、居住时间以及社会地位等是决定他们是否使用信用卡的关键因素；此外，Carow 等(1999) 认为年龄较小、受教育水平较高的以及拥有更多的各类卡种的人更有可能使用信用卡，而往往他们的信用风险也很高。还有很多学者研究了信用卡信用风险的影响

因素以及风险的可预测性，比如 Altman 等(1998)通过计算负债/收入的比率以及花费大于收入的消费者比率的方法，发现贫穷家庭更容易陷入信用卡债务的困境中，也就是说，低收入者信用卡欠款超过了他们的收入。Schreiner(2004)通过对玻利维亚的数据研究发现，女性的信用违约风险常常低于男性，这说明信用卡信用风险存在性别差异。Dinh 等(2007)用越南的数据研究发现，越南的主要人群学历较低，但学历变量对信用风险没有显著影响，高学历的违约率并不一定低，如大学毕业的群体拥有最高的信用违约风险。此外 Gross 等(2002)研究发现职业为零售业的客户信用违约风险常常高于其它的工作。

另外，还有一些学者从研究方法上的创新，运用更为高效精确的模型研究，比如 Bellotti 等(2009)利用数据挖掘的支持向量机(SVM)方法建立了信用评分模型，认为有房者的违约风险较低，且过去 6 个月申请贷款的次数越多，风险越大，同时年龄变量对此也有重要影响。Chen 等(2009)用 Hybrid SVM 对中国某银行的数据进行研究，发现对信用评分影响最大的收入变量，其次是受教育水平等，年龄变量也有一定的影响，而婚姻情况、国别、籍贯的影响则很小。Lee(2007)构建了基于支持向量机的信用风险评估模型，对违约率进行了预测。Jon 等(2008)利用自组织映射方法(SOM)研究了信用卡欺诈识别。Yu 等(2010)提出了基于多代理组合学习法的四阶段支持向量机(SVM)法研究信用卡风险评估，Twala(2010)用组合学习算法评估信用风险，研究表明该方法的准确率较之单个分类器有着显著的提高。我国学者迟国泰等(2006)建立了个人信用卡信用风险评价指标体系，利用层次分析法计算指标的权重，并建立信用评分模型。刘闽和林成德(2005)基于支持向量机信用风险评估模型实证分析了我国商业银行的风险。郭英见等(2009)利用 BP 神经网络、支持向量机和 DS 证据理论等建立了信用风险评估模型，认为该模型相对于传统神经网络、支持向量机评估模型更为有效。

1.2.2 个人住房贷款研究现状

目前学术界普遍认同将个人住房贷款信用风险研究划分为三个阶段：(20 世纪 70 年代以前)为第一阶段，根据贷款特征和贷款人特征对带违约进行评估；(20 世纪 70-80)为第二阶段引入期权理论，把违约看成是看跌期权；(20 世纪 80 年代以来)为第三阶段，研究群体的违约率。本文对近期国内外的主要研究成果进行了归纳，主要按照研究的影响因素来进行梳理。

许多学者针对住房按揭贷款违约风险的影响因素进行了研究，主要包括